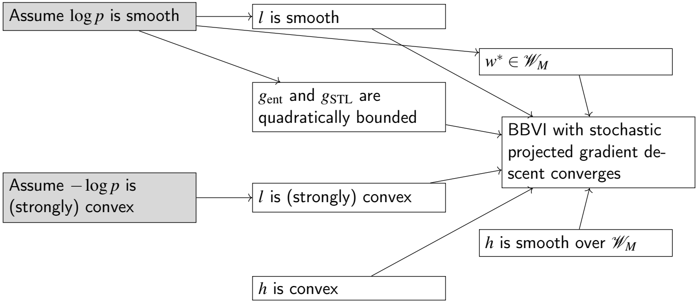
Convergence Guarantees for Variational Inference

Justin Domke, University of Massachusetts Amherst

these slides: `t.ly/sICHy` or `people.cs.umass.edu/domke/convergence.pdf`



Outline

1. Introduction
2. The neg-entropy
3. The energy
4. Proximal gradient descent
5. Projected gradient descent
6. Gradient variance
7. Real convergence guarantees
8. Discussion

Inference: Given $p(z,x)$ and observed data x, approximate $p(z|x)$

Inference: Given $p(z,x)$ and observed data x, approximate $p(z|x)$

Variational inference: ...by choosing some family $q_w(z)$ and minimizing $KL(q_w(z) \| p(z|x))$

Inference: Given $p(z,x)$ and observed data x, approximate $p(z|x)$

Variational inference: ...by choosing some family $q_w(z)$ and minimizing $KL(q_w(z)\|p(z|x))$

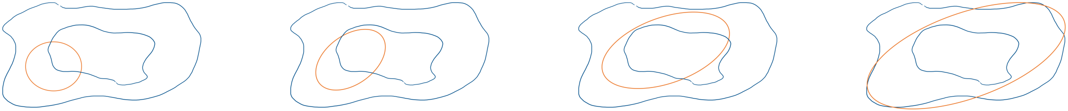
Black box variational inference: ...while only evaluating $\log p(z,x)$ or $\nabla_z \log p(z,x)$.

Black box VI in practice

- Let $q_w(z)$ be the set of dense Gaussians
- Inialize w somehow.
- Repeat:
  - Get stochastic estimate g of $\nabla_w KL\left(q_w(z) \| p(z|x)\right)$.
  - Take gradient step: $w \leftarrow w - \gamma g$.

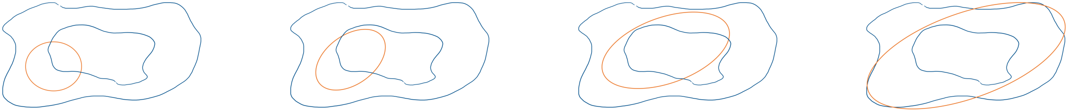
Black box VI in practice

- Let $q_w(z)$ be the set of dense Gaussians
- Inialize w somehow.
- Repeat:
  - Get stochastic estimate g of $\nabla_w KL(q_w(z) \| p(z|x))$.
  - Take gradient step: $w \leftarrow w - \gamma g$.



Black box VI in practice

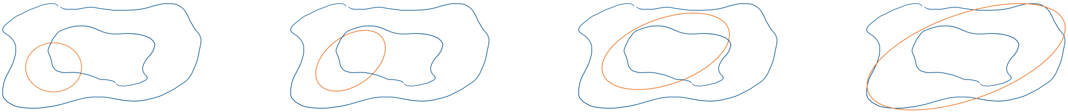
- Let $q_w(z)$ be the set of dense Gaussians
- Inialize w somehow.
- Repeat:
  - Get stochastic estimate g of $\nabla_w KL(q_w(z) \| p(z|x))$.
  - Take gradient step: $w \leftarrow w - \gamma g$.



Easy to find g via autodiff, seems to work well in practice.

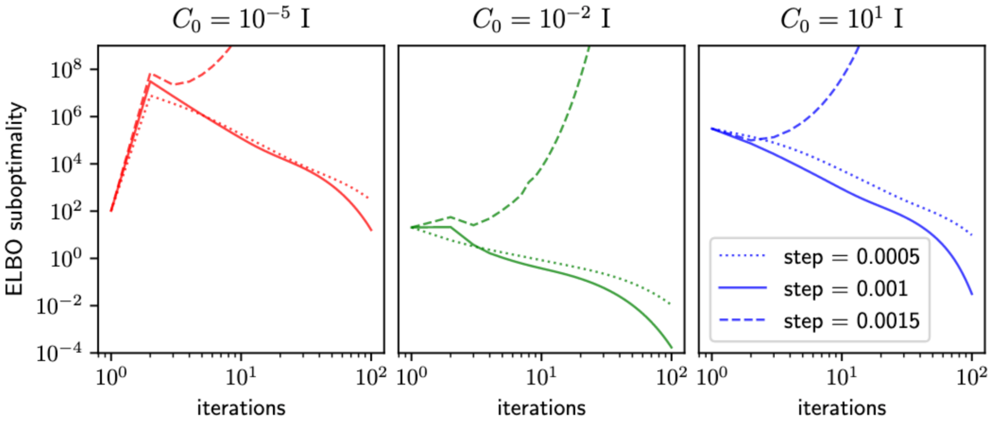
Black box VI in practice

- Let $q_w(z)$ be the set of dense Gaussians
- Inialize w somehow.
- Repeat:
  - Get stochastic estimate g of $\nabla_w KL\left(q_w(z) \| p(z|x)\right)$.
  - Take gradient step: $w \leftarrow w - \gamma g$.



Easy to find g via autodiff, seems to work well in practice.

This talk: But can we prove anything?



Results on `fires` with *exact* gradients, initialized with $\Sigma = C_0 C_0^\top$.

Bad news

Can we guarantee anything? If $p(z,x)$ could be *anything*, then no.

Bad news

Can we guarantee anything? If $p(z,x)$ could be *anything*, then no.

Best we can hope for: If p is "nice" then BBVI optimization is "nice".

How p might be nice

Plausible properties for $f(z) = -\log p(z,x)$:

How p might be nice

Plausible properties for $f(z) = -\log p(z,x)$:

- Convex ($\nabla_z^2 f(z) \succeq 0$)
- Strongly convex ($\nabla_z^2 f(z) \succeq cI$)
- Smooth ($\nabla_z^2 f(z) \preceq MI$)

How p might be nice

Plausible properties for $f(z) = -\log p(z,x)$:

- Convex ($\nabla_z^2 f(z) \succeq 0$)
- Strongly convex ($\nabla_z^2 f(z) \succeq cI$)
- Smooth ($\nabla_z^2 f(z) \preceq MI$)

$p(z,x)$	convex	strongly covex	smooth
Gaussian	✓	✓	✓
Bayesian linear regression	✓	✓	✓
Bayesian logistic regression	✓	✓	✓
Heirarchical logistic regression	✓	×	✓

But is optimization nice?

$$\min_w F(w) := \underbrace{\mathop{\mathbb{E}}_{q_w(z)} [-\log p(z,x)]}_{\text{"energy" } l(w)} + \underbrace{\mathop{\mathbb{E}}_{q_w(z)} \log q_w(z)}_{\text{"neg-entropy" } h(w)}$$

But is optimization nice?

$$\min_{w} F(w) := \underbrace{\mathbb{E}_{q_w(z)}[-\log p(z,x)]}_{\text{"energy" } l(w)} + \underbrace{\mathbb{E}_{q_w(z)} \log q_w(z)}_{\text{"neg-entropy" } h(w)}$$

Assume henceforth that $q_w(z) = \mathcal{N}(z|m, CC^\top), \quad w = (m, C)$.

But is optimization nice?

$$\min_w F(w) := \underbrace{\mathbb{E}_{q_w(z)}[-\log p(z,x)]}_{\text{"energy" } l(w)} + \underbrace{\mathbb{E}_{q_w(z)} \log q_w(z)}_{\text{"neg-entropy" } h(w)}$$

Assume henceforth that $q_w(z) = \mathcal{N}(z|m, CC^\top), \quad w = (m, C)$.

How stochastic optimization guarantees usually work:

1. Prove that gradient has **bounded noise** (either $\mathbb{E}\|g\|_2^2 \leq b$ or $\mathbb{V}[g] \leq b$)
2. Prove that objective is **convex** or **strongly convex**
3. Prove that objective is Lipschitz **smooth**.

But is optimization nice?

$$\min_w F(w) := \underbrace{\mathop{\mathbb{E}}_{q_w(z)}[-\log p(z,x)]}_{\text{"energy" } l(w)} + \underbrace{\mathop{\mathbb{E}}_{q_w(z)} \log q_w(z)}_{\text{"neg-entropy" } h(w)}$$

Assume henceforth that $q_w(z) = \mathcal{N}(z|m, CC^\top), \quad w = (m, C)$.

How stochastic optimization guarantees usually work:

1. Prove that gradient has **bounded noise** (either $\mathbb{E}\|g\|_2^2 \leq b$ or $\mathbb{V}[g] \leq b$)
2. Prove that objective is **convex** or **strongly convex**
3. Prove that objective is Lipschitz **smooth**.

Trouble: If $p(z|x) = \mathcal{N}(z|0, I)$, then 1 and 3 are false!

Table of properties

$$F(w) := \underbrace{\mathbb{E}_{q_w(z)}[-\log p(z,x)]}_{\text{"energy" } l(w)} + \underbrace{\mathbb{E}_{q_w(z)} \log q_w(z)}_{\text{"neg-entropy" } h(w)}$$

Condition on $-\log p(z,x)$	**Consequence**
none	
convex	
c-strongly convex	
M-smooth	

Outline

1. Introduction
2. The neg-entropy
3. The energy
4. Proximal gradient descent
5. Projected gradient descent
6. Gradient variance
7. Real convergence guarantees
8. Discussion

Neg-entropy

Theorem

$h(w)$ *is convex, but not strongly convex and not smooth.*

Neg-entropy

Theorem

$h(w)$ is convex, but not strongly convex and not smooth.

Proof.

$h(w) = -\log|\det C| + \frac{d}{2}\log(2\pi e)$

□

Neg-entropy

Theorem

$h(w)$ is convex, but not strongly convex and not smooth.

Proof.

$h(w) = -\log|\det C| + \frac{d}{2}\log(2\pi e) = -\sum_i \log\sigma_i(C) + \text{const.}$ □

Blows up when singular values of C are small.

Table of properties

$$F(w) := \underbrace{\mathbb{E}_{q_w(z)}[-\log p(z,x)]}_{\text{"energy" } l(w)} + \underbrace{\mathbb{E}_{q_w(z)} \log q_w(z)}_{\text{"neg-entropy" } h(w)}$$

Condition on $-\log p(z,x)$	**Consequence**
none	$h(w)$ convex (when C symmetric or triangular) $h(w)$ *not* strongly convex, *not* smooth
convex	
c-strongly convex	
M-smooth	

Outline

1. Introduction
2. The neg-entropy
3. The energy
4. Proximal gradient descent
5. Projected gradient descent
6. Gradient variance
7. Real convergence guarantees
8. Discussion

(Strong) convexity

Theorem

If $-\log p(z,x)$ *is convex, then* $l(w)$ *is also convex.*

Theorem

If $-\log p(z,x)$ *is* c*-strongly convex, then* $l(w)$ *is also* c*-strongly convex.*

(Strong) convexity

Theorem

If $-\log p(z,x)$ *is convex, then* $l(w)$ *is also convex.*

Theorem

If $-\log p(z,x)$ *is* c*-strongly convex, then* $l(w)$ *is also* c*-strongly convex.*

Proof.

Easy.
(Convexity result due to Titsias and Lázaro-gredilla (2014))
(Strong convexity result (D., 2019) generalizes Challis and Barber (2013)) □

Smoothness

Theorem

If $\log p(z,x)$ *is* M*-smooth, then* $l(w)$ *is also* M*-smooth.*

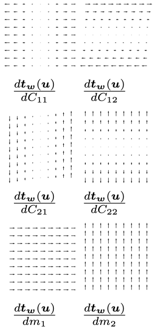
Smoothness

Theorem

If $\log p(z,x)$ is M-smooth, then $l(w)$ is also M-smooth.

Proof.

Define inner-product space + Bessel's inequality + various exact calculations.



Lemma 2. *$\langle \boldsymbol{a}, \boldsymbol{b} \rangle_s = \mathbb{E}_{u\sim s}\, \boldsymbol{a}(u)^\top \boldsymbol{b}(u)$ is a valid inner-product on squared-integrable $\boldsymbol{a} : \mathbb{R}^d \to \mathbb{R}^k$.*

Proof. The space of square integrable functions is $\left\{\boldsymbol{a} : \mathbb{R}^d \to \mathbb{R}^k \mid \mathbb{E}_{u\sim s}\, a_i(u)^2 \leq \infty \, \forall i \in \{1, ..., k\}\right\}$. Since each component $a_i(\boldsymbol{u})$ and $b_i(\boldsymbol{u})$ is square-integrable with respect to $s(\boldsymbol{u})$ we know (by Cauchy-Schwarz) that $\mathbb{E}_{u\sim s}\, a_i(u)b_i(u) \leq \sqrt{\mathbb{E}_{u\sim s}\, a_i(u)^2}\sqrt{\mathbb{E}_{u\sim s}\, b_i(u)}$ is finite and real. Therefore, we have by linearity of expectation that

$$\begin{aligned}\sum_{i=1}^k \mathop{\mathbb{E}}_{u\sim s} a_i(u)b_i(u) &= \mathop{\mathbb{E}}_{u\sim s} \sum_{i=1}^k a_i(u)b_i(u) \\ &= \mathop{\mathbb{E}}_{u\sim s} \boldsymbol{a}(u)^\top \boldsymbol{b}(u) \\ &= \langle \boldsymbol{a}, \boldsymbol{b} \rangle_s\end{aligned}$$

is finite and real for all $\boldsymbol{a}, \boldsymbol{b} \in V_s$. To show that $(V_s, \langle \cdot, \cdot \rangle_s)$ is a valid inner-product space, it is easy to establish all the necessary properties of the inner-product, namely for all $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in V_s$,

$\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \langle \boldsymbol{b}, \boldsymbol{a} \rangle$

$\langle \theta \boldsymbol{a}, \boldsymbol{b} \rangle = \theta \langle \boldsymbol{a}, \boldsymbol{b} \rangle$ for $\theta \in \mathbb{R}$

$\langle \boldsymbol{a} + \boldsymbol{b}, \boldsymbol{c} \rangle = \langle \boldsymbol{a}, \boldsymbol{c} \rangle + \langle \boldsymbol{b}, \boldsymbol{c} \rangle$

$\langle \boldsymbol{a}, \boldsymbol{a} \rangle \geq 0$

$\langle \boldsymbol{a}, \boldsymbol{a} \rangle = 0 \Leftrightarrow \boldsymbol{a} = \boldsymbol{0}$. (Where $\boldsymbol{0}(\boldsymbol{\varepsilon})$ is a function that always returns a vector of k zeros.) □

Lemma 3. *Let $\boldsymbol{a}_i(\boldsymbol{u}) = \frac{d}{dw_i}\boldsymbol{t}_{\boldsymbol{w}}(\boldsymbol{u})$. This is independent of $\boldsymbol{w}$ and $\frac{dl(\boldsymbol{w})}{dw_i} = \langle \boldsymbol{a}_i, \nabla f \circ \boldsymbol{t}_{\boldsymbol{w}} \rangle_s$.*

Proof. Now, we can write $l(\boldsymbol{w})$ as

$$l(\boldsymbol{w}) = \mathop{\mathbb{E}}_{\mathbf{z}\sim q_{\boldsymbol{w}}} f(\mathbf{z}) = \mathop{\mathbb{E}}_{u\sim s} f(\boldsymbol{t}_{\boldsymbol{w}}(u)).$$

Since $\boldsymbol{t}_{\boldsymbol{w}}(\boldsymbol{u}) = C\boldsymbol{u} + \boldsymbol{m}$ is an affine function, it's easy to see that both $\frac{d}{dC_{ij}}\boldsymbol{t}_{\boldsymbol{w}}(\boldsymbol{u})$ and $\frac{d}{dm_i}\boldsymbol{t}_{\boldsymbol{w}}(\boldsymbol{u})$ are independent of $\boldsymbol{w}$. Therefore, the gradient of $l(\boldsymbol{w})$ can be written as

$$\begin{aligned}\nabla_{w_i} l(\boldsymbol{w}) &= \nabla_{w_i} \mathop{\mathbb{E}}_{u\sim s} f(\boldsymbol{t}_{\boldsymbol{w}}(u)) \\ &= \mathop{\mathbb{E}}_{u\sim s} \nabla_{w_i}\boldsymbol{t}_{\boldsymbol{w}}(u)^\top \nabla f(\boldsymbol{t}_{\boldsymbol{w}}(u)). \\ &= \langle \boldsymbol{a}_i, \nabla f \circ \boldsymbol{t}_{\boldsymbol{w}} \rangle_s.\end{aligned}$$

□

Lemma 4. *If s is standardized, then the functions $\{\boldsymbol{a}_i\}$ are orthonormal in $\langle \cdot, \cdot \rangle_s$.*

Proof. It is easy to calculate that

$$\begin{aligned}\frac{d}{dm_i}\boldsymbol{t}_{\boldsymbol{w}}(\boldsymbol{u}) &= \boldsymbol{e}_i \\ \frac{d}{dC_{ij}}\boldsymbol{t}_{\boldsymbol{w}}(\boldsymbol{u}) &= \boldsymbol{e}_i u_j,\end{aligned}$$

where $\boldsymbol{e}_i$ is the indicator vector in the i-th component. Therefore, we have that

$$\begin{aligned}&\mathop{\mathbb{E}}_{u\sim s}\left(\frac{d}{dm_i}\boldsymbol{t}_{\boldsymbol{w}}(u)\right)^\top\left(\frac{d}{dm_j}\boldsymbol{t}_{\boldsymbol{w}}(u)\right) \\ &= \mathop{\mathbb{E}}_{u\sim s} \boldsymbol{e}_i^\top \boldsymbol{e}_j \\ &= I[i=j] \\ &\mathop{\mathbb{E}}_{u\sim s}\left(\frac{d}{dC_{ij}}\boldsymbol{t}_{\boldsymbol{w}}(u)\right)^\top\left(\frac{d}{dm_k}\boldsymbol{t}_{\boldsymbol{w}}(u)\right) \\ &= \mathop{\mathbb{E}}_{u\sim s} u_j \boldsymbol{e}_i^\top \boldsymbol{e}_k \\ &= I[i=k] \mathop{\mathbb{E}}_{u\sim s} u_j \\ &= 0 \\ &\text{(since zero mean)} \\ &\mathop{\mathbb{E}}_{u\sim s}\left(\frac{d}{dC_{ij}}\boldsymbol{t}_{\boldsymbol{w}}(u)\right)^\top\left(\frac{d}{dC_{kl}}\boldsymbol{t}_{\boldsymbol{w}}(u)\right) \\ &= \mathbb{E}\, u_j u_l \boldsymbol{e}_i^\top \boldsymbol{e}_k \\ &= I[i=k] \mathop{\mathbb{E}}_{u\sim s} u_j u_l \\ &= I[i=k] I[j=l] \\ &\text{(since unit variance and zero mean)}\end{aligned}$$

These three identities are equivalent to stating that $\{\boldsymbol{a}_i\}$ are orthonormal in $\langle \cdot, \cdot \rangle_s$. □

Lemma 5. *If s is standardized, then $\mathbb{E}_{u\sim s} \|\boldsymbol{t}_{\boldsymbol{w}}(u) - \boldsymbol{t}_{\boldsymbol{v}}(u)\|_2^2 = \|\boldsymbol{w} - \boldsymbol{v}\|_2^2$.*

Proof. Let $\Delta \boldsymbol{m}$ and ΔS denote the difference of the $\boldsymbol{m}$ and S parts of $\boldsymbol{w}$, respectively. We want to calculate

$$\begin{aligned}&\mathop{\mathbb{E}}_{u\sim s} \|\boldsymbol{t}_{\boldsymbol{w}}(u) - \boldsymbol{t}_{\boldsymbol{v}}(u)\|_2^2 \\ &= \mathop{\mathbb{E}}_{u\sim s} \|\Delta C \boldsymbol{\varepsilon} + \Delta \boldsymbol{m}\|_2^2 \\ &= \mathop{\mathbb{E}}_{u\sim s} \left(\|(\Delta C)u\|_2^2 + 2\Delta \boldsymbol{m}^\top \Delta C u + \|\Delta \boldsymbol{m}\|_2^2\right).\end{aligned}$$

It is easy to see that the expectation of the middle term is zero, and the last is a constant. The expectation of the first term is

$$\begin{aligned}\mathop{\mathbb{E}}_{u\sim s} \|(\Delta C)u\|_2^2 &= \mathop{\mathbb{E}}_{u\sim s} u^\top (\Delta C)^\top (\Delta C) u \\ &= \mathop{\mathbb{E}}_{u\sim s} \operatorname{tr}\left(u^\top (\Delta C)^\top (\Delta C) u\right) \\ &= \mathop{\mathbb{E}}_{u\sim s} \operatorname{tr}\left((\Delta C)^\top (\Delta C) u u^\top\right) \\ &= \operatorname{tr}\left((\Delta C)^\top (\Delta C)\right) = \|\nabla C\|_F^2. \\ &\quad \text{(since zero mean and unit variance)}\end{aligned}$$

Putting this together gives that

$$\begin{aligned}\mathop{\mathbb{E}}_{u\sim s} \|\boldsymbol{t}_{\boldsymbol{w}}(u) - \boldsymbol{t}_{\boldsymbol{v}}(u)\|_2^2 &= \|\Delta C\|_F^2 + \|\Delta \boldsymbol{m}\|_2^2 \\ &= \|\boldsymbol{w} - \boldsymbol{v}\|_2^2.\end{aligned}$$

□

Proof of Thm. 1. Take two parameter vectors, $\boldsymbol{w}$ and $\boldsymbol{v}$. Apply Lem. 3 to each component of the gradients $\nabla l(\boldsymbol{w})$ and $\nabla l(\boldsymbol{v})$ to get that

$$\begin{aligned}&\|\nabla l(\boldsymbol{w}) - \nabla l(\boldsymbol{v})\|_2^2 \\ &= \sum_i \left(\langle \boldsymbol{a}_i, \nabla f \circ \boldsymbol{t}_{\boldsymbol{w}} \rangle_s - \langle \boldsymbol{a}_i, \nabla f \circ \boldsymbol{t}_{\boldsymbol{v}} \rangle_s\right)^2 \\ &= \sum_i \langle \boldsymbol{a}_i, \nabla f \circ \boldsymbol{t}_{\boldsymbol{w}} - \nabla f \circ \boldsymbol{t}_{\boldsymbol{v}} \rangle_s^2.\end{aligned}$$

Lem. 4 showed that the functions $\{\boldsymbol{a}_i\}$ are orthonormal in the inner-product $\langle \cdot, \cdot \rangle_s$. Thus, by Bessel's inequality,

$$\begin{aligned}\|\nabla l(\boldsymbol{w}) - \nabla l(\boldsymbol{v})\|_2^2 &\leq \|\nabla f \circ \boldsymbol{t}_{\boldsymbol{w}} - \nabla f \circ \boldsymbol{t}_{\boldsymbol{v}}\|_s^2, \qquad (5) \\ &= \mathop{\mathbb{E}}_{u\sim s} \|\nabla f(\boldsymbol{t}_{\boldsymbol{w}}(u)) - \nabla f(\boldsymbol{t}_{\boldsymbol{v}}(u))\|_2^2\end{aligned}$$

where $\|\cdot\|_s$ denotes the norm corresponding to $\langle \cdot, \cdot \rangle_s$. Now apply the smoothness of f to get that

$$\begin{aligned}\|\nabla l(\boldsymbol{w}) - \nabla l(\boldsymbol{v})\|_2^2 &\leq M^2 \mathop{\mathbb{E}}_{u\sim s} \|\boldsymbol{t}_{\boldsymbol{w}}(u) - \boldsymbol{t}_{\boldsymbol{v}}(u)\|_2^2 \qquad (6) \\ &= M^2 \|\boldsymbol{w} - \boldsymbol{v}\|_2^2, \qquad (7)\end{aligned}$$

where the last equality follows from Lem. 5. □

□

Table of properties

$$F(w) := \underbrace{\mathbb{E}_{q_w(z)}[-\log p(z,x)]}_{\text{"energy" } l(w)} + \underbrace{\mathbb{E}_{q_w(z)} \log q_w(z)}_{\text{"neg-entropy" } h(w)}$$

Condition on $-\log p(z,x)$	**Consequence**
none	$h(w)$ convex (when C symmetric or triangular) $h(w)$ *not* strongly convex, *not* smooth
convex	$l(w)$ convex
c-strongly convex	$l(w)$ c-strongly convex
M-smooth	$l(w)$ M-smooth

Outline

1. Introduction
2. The neg-entropy
3. The energy
4. Proximal gradient descent
5. Projected gradient descent
6. Gradient variance
7. Real convergence guarantees
8. Discussion

Challenge: Non-smooth objective

$$F(w) := \underbrace{\mathop{\mathbb{E}}_{q_w(z)}[-\log p(z,x)]}_{\text{"energy" } l(w)} + \underbrace{\mathop{\mathbb{E}}_{q_w(z)} \log q_w(z)}_{\text{"neg-entropy" } h(w)}$$

Problem: h is not smooth. So F is (probably) not smooth.

Gradient descent (need $l+h$ smooth)

$$
\begin{aligned}
w' &= w - \gamma(\nabla l(w) + \nabla h(w)) \\
&= \operatorname*{argmin}_v \underbrace{l(w) + h(w) + \langle \nabla l(w) + \nabla h(w),\ v - w\rangle}_{\text{local affine approximation of } l(v)+h(v)} + \underbrace{\frac{1}{2\gamma}\|v-w\|_2^2}_{\text{penalty term}}
\end{aligned}
$$

Gradient descent (need $l+h$ smooth)

$$
\begin{aligned}
w' &= w - \gamma(\nabla l(w) + \nabla h(w)) \\
&= \operatorname*{argmin}_v \underbrace{l(w) + h(w) + \langle \nabla l(w) + \nabla h(w),\ v - w\rangle}_{\text{local affine approximation of } l(v)+h(v)} + \underbrace{\frac{1}{2\gamma}\|v-w\|_2^2}_{\text{penalty term}}
\end{aligned}
$$

Proximal gradient descent: (only need l smooth)

$$
\begin{aligned}
w' &= \operatorname*{argmin}_v \ \underbrace{l(w) + \langle \nabla l(w),\ v - w\rangle}_{\text{local affine approximation of } l} + \underbrace{h(v)}_{\text{exact } h} + \underbrace{\frac{1}{2\gamma}\|v-w\|_2^2}_{\text{penalty term}} \\
&= \operatorname*{prox}_{\gamma h}[w - \gamma \nabla l(w)]
\end{aligned}
$$

Gradient descent (need $l+h$ smooth)

$$
\begin{aligned}
w' &= w - \gamma(\nabla l(w) + \nabla h(w)) \\
&= \operatorname*{argmin}_v \underbrace{l(w) + h(w) + \langle \nabla l(w) + \nabla h(w),\ v - w\rangle}_{\text{local affine approximation of } l(v)+h(v)} + \underbrace{\frac{1}{2\gamma}\|v-w\|_2^2}_{\text{penalty term}}
\end{aligned}
$$

Proximal gradient descent: (only need l smooth)

$$
\begin{aligned}
w' &= \operatorname*{argmin}_v \underbrace{l(w) + \langle \nabla l(w),\ v - w\rangle}_{\text{local affine approximation of } l} + \underbrace{h(v)}_{\text{exact } h} + \underbrace{\frac{1}{2\gamma}\|v-w\|_2^2}_{\text{penalty term}} \\
&= \operatorname*{prox}_{\gamma h}[w - \gamma\nabla l(w)]
\end{aligned}
$$

Computing $\mathrm{prox}_{\gamma h}[w] = \mathrm{argmin}_v\, h(v) + \frac{1}{2\gamma}\|w-v\|_2^2$ is easy when C is triangular.

Gradient descent (need $l+h$ smooth)

$$\begin{aligned} w' &= w - \gamma(\nabla l(w) + \nabla h(w)) \\ &= \operatorname*{argmin}_v \underbrace{l(w) + h(w) + \langle \nabla l(w) + \nabla h(w),\ v - w\rangle}_{\text{local affine approximation of } l(v)+h(v)} + \underbrace{\frac{1}{2\gamma}\|v-w\|_2^2}_{\text{penalty term}} \end{aligned}$$

Proximal gradient descent: (only need l smooth)

$$\begin{aligned} w' &= \operatorname*{argmin}_v \underbrace{l(w) + \langle \nabla l(w),\ v - w\rangle}_{\text{local affine approximation of } l} + \underbrace{h(v)}_{\text{exact } h} + \underbrace{\frac{1}{2\gamma}\|v-w\|_2^2}_{\text{penalty term}} \\ &= \operatorname*{prox}_{\gamma h}[w - \gamma \nabla l(w)] \end{aligned}$$

Computing $\operatorname{prox}_{\gamma h}[w] = \operatorname{argmin}_v h(v) + \frac{1}{2\gamma}\|w-v\|_2^2$ is easy when C is triangular.

Standard theory: Converges if l is (strongly) convex and smooth.

Outline

1. Introduction
2. The neg-entropy
3. The energy
4. Proximal gradient descent
5. **Projected gradient descent**
6. Gradient variance
7. Real convergence guarantees
8. Discussion

Solution guarantees

$$F(w) := \underbrace{\mathop{\mathbb{E}}_{q_w(z)}[-\log p(z,x)]}_{\text{"energy" } l(w)} + \underbrace{\mathop{\mathbb{E}}_{q_w(z)} \log q_w(z)}_{\text{"neg-entropy" } h(w)}$$

Hmmmm...

Solution guarantees

$$F(w) := \underbrace{\mathbb{E}_{q_w(z)}[-\log p(z,x)]}_{\text{"energy" } l(w)} + \underbrace{\mathbb{E}_{q_w(z)} \log q_w(z)}_{\text{"neg-entropy" } h(w)}$$

Hmmmm...

- $h(w) = -\log|\det C| +$ const. is smooth except when singular values of C are small.

Solution guarantees

$$F(w) := \underbrace{\mathop{\mathbb{E}}_{q_w(z)}[-\log p(z,x)]}_{\text{"energy" } l(w)} + \underbrace{\mathop{\mathbb{E}}_{q_w(z)} \log q_w(z)}_{\text{"neg-entropy" } h(w)}$$

Hmmmm...

- $h(w) = -\log|\det C| + \text{const.}$ is smooth except when singular values of C are small.
- $h(w)$ also becomes really *large* when the singular values of C are small.

Solution guarantees

$$F(w) := \underbrace{\mathop{\mathbb{E}}_{q_w(z)} [-\log p(z,x)]}_{\text{"energy" } l(w)} + \underbrace{\mathop{\mathbb{E}}_{q_w(z)} \log q_w(z)}_{\text{"neg-entropy" } h(w)}$$

Hmmmm...

- $h(w) = -\log|\det C| + \text{const.}$ is smooth except when singular values of C are small.
- $h(w)$ also becomes really *large* when the singular values of C are small.
- Maybe the singular values of C can't be too small *at the solution*?

Solution guarantees

$$F(w) := \underbrace{\mathop{\mathbb{E}}_{q_w(z)} [-\log p(z,x)]}_{\text{"energy" } l(w)} + \underbrace{\mathop{\mathbb{E}}_{q_w(z)} \log q_w(z)}_{\text{"neg-entropy" } h(w)}$$

Hmmmm...

- $h(w) = -\log|\det C| + \text{const.}$ is smooth except when singular values of C are small.
- $h(w)$ also becomes really *large* when the singular values of C are small.
- Maybe the singular values of C can't be too small *at the solution*?
- And maybe we can exploit that somehow?

$$\mathscr{W}_M := \left\{ (m, C) \middle| \sigma_{\min}(C) \geq \frac{1}{\sqrt{M}} \right\}$$

$$\mathscr{W}_M := \left\{ (m, C) | \sigma_{\min}(C) \geq \frac{1}{\sqrt{M}} \right\}$$

Theorem

If $\log p(z,x)$ *is* M*-smooth and* w^* *minimizes* $l(w) + h(w)$, *then* $w^* \in \mathscr{W}_M$. *(D. 2020, Thm. 7)*

$$\mathscr{W}_M := \left\{ (m, C) | \sigma_{\min}(C) \geq \frac{1}{\sqrt{M}} \right\}$$

Theorem

If $\log p(z,x)$ *is* M*-smooth and* w^* *minimizes* $l(w) + h(w)$, *then* $w^* \in \mathscr{W}_M$. *(D. 2020, Thm. 7)*

Lemma

$h(w)$ *is* M*-smooth over* $\mathscr{W}_{\mathscr{M}}$. *(D. 2020, Lemma 12)*

$$\mathscr{W}_M := \left\{ (m, C) | \sigma_{\min}(C) \geq \frac{1}{\sqrt{M}} \right\}$$

Theorem

If $\log p(z, x)$ *is* M*-smooth and* w^* *minimizes* $l(w) + h(w)$*, then* $w^* \in \mathscr{W}_M$*. (D. 2020, Thm. 7)*

Lemma

$h(w)$ *is* M*-smooth over* $\mathscr{W}_{\mathscr{M}}$*. (D. 2020, Lemma 12)*

Projected gradient descent:

$$w' = \text{proj}_{\mathscr{W}_M}[w - \gamma(\nabla l(w) + \nabla h(w))]$$

$\text{proj}_{\mathscr{W}_M}[w] = \text{argmin}_{w' \in \mathscr{W}_M} \|w - w'\|_2$

$$\mathscr{W}_M := \left\{(m,C) | \sigma_{\min}(C) \geq \frac{1}{\sqrt{M}}\right\}$$

Theorem

If $\log p(z,x)$ is M-smooth and w^* minimizes $l(w)+h(w)$, then $w^* \in \mathscr{W}_M$. (D. 2020, Thm. 7)

Lemma

$h(w)$ is M-smooth over $\mathscr{W}_{\mathscr{M}}$. (D. 2020, Lemma 12)

Projected gradient descent:

$$w' = \mathrm{proj}_{\mathscr{W}_M}[w - \gamma(\nabla l(w) + \nabla h(w))]$$

$\mathrm{proj}_{\mathscr{W}_M}[w] = \mathrm{argmin}_{w' \in \mathscr{W}_M} \|w - w'\|_2$ is easy to compute but requires an SVD of C.

$$\mathscr{W}_M := \left\{ (m, C) | \sigma_{\min}(C) \geq \frac{1}{\sqrt{M}} \right\}$$

Theorem

If $\log p(z,x)$ *is* M*-smooth and* w^* *minimizes* $l(w) + h(w)$*, then* $w^* \in \mathscr{W}_M$*. (D. 2020, Thm. 7)*

Lemma

$h(w)$ *is* M*-smooth over* $\mathscr{W}_\mathscr{M}$*. (D. 2020, Lemma 12)*

Projected gradient descent:

$$w' = \mathrm{proj}_{\mathscr{W}_M}[w - \gamma(\nabla l(w) + \nabla h(w))]$$

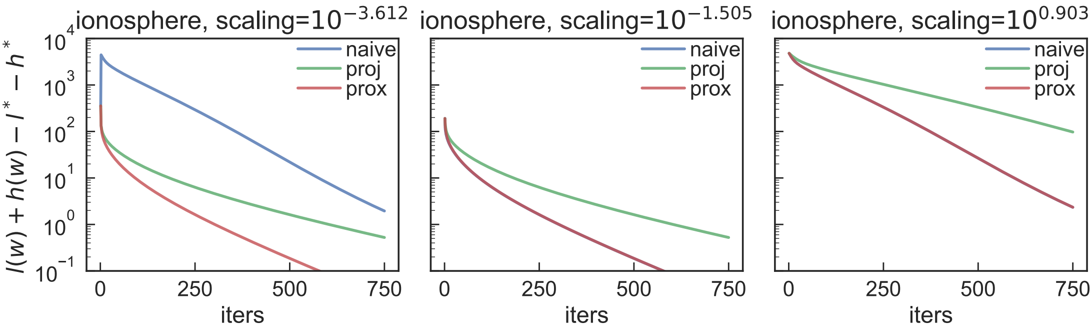
$\mathrm{proj}_{\mathscr{W}_M}[w] = \mathrm{argmin}_{w' \in \mathscr{W}_M} \|w - w'\|_2$ is easy to compute but requires an SVD of C.

Standard theory: converges if $l + h$ is (strongly) convex and smooth.

Table of properties

$$F(w) := \underbrace{\mathbb{E}_{q_w(z)}[-\log p(z,x)]}_{\text{"energy" } l(w)} + \underbrace{\mathbb{E}_{q_w(z)} \log q_w(z)}_{\text{"neg-entropy" } h(w)}$$

Condition on $-\log p(z,x)$	Consequence
none	$h(w)$ convex (when C symmetric or triangular)
	$h(w)$ *not* strongly convex, *not* smooth
	$h(w)$ is M-smooth over $\mathscr{W}_M$
convex	$l(w)$ convex
c-strongly convex	$l(w)$ c-strongly convex
M-smooth	$l(w)$ M-smooth
	$w^* \in \mathscr{W}_M$



Bayesian logistic regression. ("Exact" gradients by reducing evaluation of 1-D integral, precomputed using numerical quadrature.)

Outline

1. Introduction
2. The neg-entropy
3. The energy
4. Proximal gradient descent
5. Projected gradient descent
6. Gradient variance
7. Real convergence guarantees
8. Discussion

Summary so far

BBVI with proximal or projected gradient descent converges, assuming:

1. $-\log p(z,x)$ is smooth
2. $-\log p(z,x)$ is (strongly) convex
3. You can compute the exact gradient.

Summary so far

BBVI with proximal or projected gradient descent converges, assuming:

1. $-\log p(z,x)$ is smooth $\longleftarrow$ Sometimes true
2. $-\log p(z,x)$ is (strongly) convex $\longleftarrow$ Sometimes true
3. You can compute the exact gradient. $\longleftarrow$ Almost never true

Estimating gradients

Can "reparameterize" using $t_w(u) = Cu + m$:

$$l(w) = -\mathop{\mathbb{E}}_{q_w(z)} \log p(z,x) = -\mathop{\mathbb{E}}_{\mathcal{N}(u|0,I)} \log p(t_w(u),x).$$

Estimating gradients

Can "reparameterize" using $t_w(u) = Cu + m$:

$$l(w) = -\mathop{\mathbb{E}}_{q_w(z)} \log p(z, x) = -\mathop{\mathbb{E}}_{\mathcal{N}(u|0,I)} \log p(t_w(u), x).$$

Definition

Typical gradient estimator (for $\nabla l(w)$):

$$g_{\text{energy}} = -\nabla_w \log p\left(t_w(u), x\right)$$

Estimating gradients

Can "reparameterize" using $t_w(u) = Cu + m$:

$$l(w) = -\mathop{\mathbb{E}}_{q_w(z)} \log p(z,x) = -\mathop{\mathbb{E}}_{\mathcal{N}(u|0,I)} \log p(t_w(u),x).$$

Definition

Typical gradient estimator (for $\nabla l(w)$):

$$g_{\text{energy}} = -\nabla_w \log p\left(t_w(u),x\right)$$

Definition

Other gradient estimators (for $\nabla l(w) + \nabla h(w)$):

$$\begin{aligned} g_{\text{ent}} &= -\nabla_w \log p\left(t_w(u),x\right) + \nabla_w h(w) \\ g_{\text{STL}} &= -\nabla_w \log p\left(t_w(u),x\right) + \left[\nabla_w \log q_v(t_w(u))\right]_{v=w} \end{aligned}$$

Quadratic bounds

Stochastic optimization proofs often assume $\mathbb{E}\,\|g\|_2^2$ (or $\mathbb{V}[g]$) is *uniformly* bounded. Not true for us!

Quadratic bounds

Stochastic optimization proofs often assume $\mathbb{E}\|g\|_2^2$ (or $\mathbb{V}[g]$) is *uniformly* bounded. Not true for us!

Definition

A gradient estimator g for $\nabla\phi$ is **quadratically bounded** with parameters (a, b, w^*) if $\mathbb{E}[g] = \nabla\phi(w)$ and

$$\mathbb{E}\|g\|_2^2 \leq a\|w - w^*\|_2^2 + b.$$

Quadratic bounds

Stochastic optimization proofs often assume $\mathbb{E}\|g\|_2^2$ (or $\mathbb{V}[g]$) is *uniformly* bounded. Not true for us!

Definition

A gradient estimator g for $\nabla\phi$ is **quadratically bounded** with parameters (a, b, w^*) if $\mathbb{E}[g] = \nabla\phi(w)$ and

$$\mathbb{E}\|g\|_2^2 \leq a\|w - w^*\|_2^2 + b.$$

Theorem

If $\log p(z, x)$ *is* M*-smooth, then* g_{energy}, g_{ent}, *and* g_{STL} *are all quadratically bounded (D., 2019, D., Garrigos, and Gower, 2023)*

Table of properties

$$F(w) := \underbrace{\mathbb{E}_{q_w(z)} [-\log p(z,x)]}_{\text{"energy" } l(w)} + \underbrace{\mathbb{E}_{q_w(z)} \log q_w(z)}_{\text{"neg-entropy" } h(w)}$$

Condition on $-\log p(z,x)$	**Consequence**
none	$h(w)$ convex (when C symmetric or triangular)
	$h(w)$ *not* strongly convex, *not* smooth
	$h(w)$ is M-smooth over $\mathscr{W}_M$
convex	$l(w)$ convex
c-strongly convex	$l(w)$ c-strongly convex
M-smooth	$l(w)$ M-smooth
	$w^* \in \mathscr{W}_M$
	gradient estimators quadratically bounded

Outline

1. Introduction
2. The neg-entropy
3. The energy
4. Proximal gradient descent
5. Projected gradient descent
6. Gradient variance
7. Real convergence guarantees
8. Discussion

An optimization "hole"

We have:

- Varying noise (quadratically bounded).
- Composite non-smooth objective.
- Objective is smooth inside of $\mathscr{W}_M$, but not *locally* smooth.

An optimization "hole"

We have:

- Varying noise (quadratically bounded).
- Composite non-smooth objective.
- Objective is smooth inside of $\mathscr{W}_M$, but not *locally* smooth.

Questions:

- Does proximal gradient descent work with quadratically bounded noise?
- Does projected gradient descent work with quadratically bounded noise?

New optimization theory

Does stochastic proximal gradient descent work with quadratically bounded noise?

New optimization theory

Does stochastic proximal gradient descent work with quadratically bounded noise?

Theorem

Yes. Converges at a $1/T$ rate if objective is smooth and strongly convex, or $1/\sqrt{T}$ if smooth and merely convex. (D., Gairrigos, and Gower, 2023, Thms. 7+8)

New optimization theory

Does stochastic proximal gradient descent work with quadratically bounded noise?

Theorem

Yes. Converges at a $1/T$ rate if objective is smooth and strongly convex, or $1/\sqrt{T}$ if smooth and merely convex. (D., Gairrigos, and Gower, 2023, Thms. 7+8)

Does stochastic projected gradient descent work with quadratically bounded noise?

New optimization theory

Does stochastic proximal gradient descent work with quadratically bounded noise?

Theorem

Yes. Converges at a $1/T$ *rate if objective is smooth and strongly convex, or* $1/\sqrt{T}$ *if smooth and merely convex. (D., Gairrigos, and Gower, 2023, Thms. 7+8)*

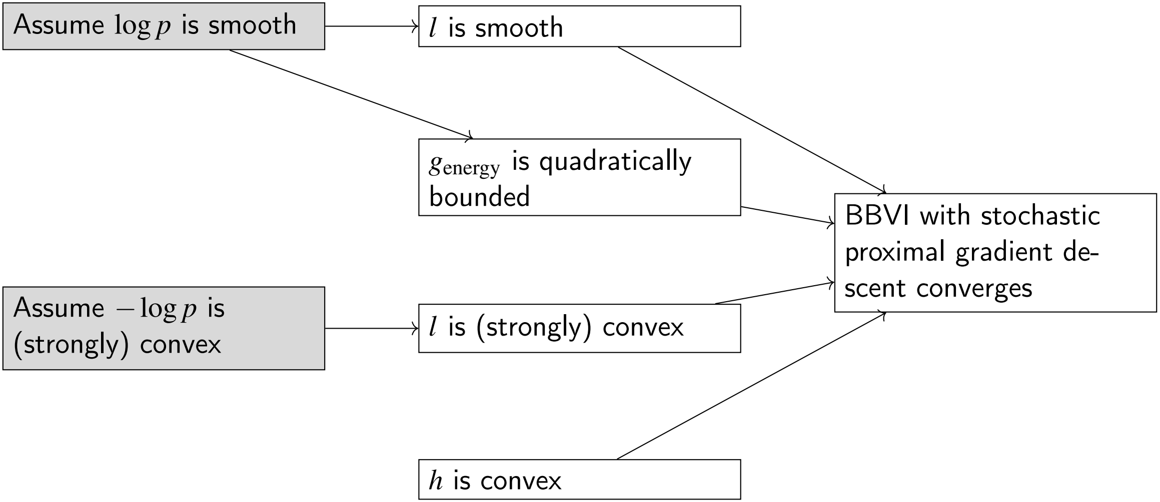
Does stochastic projected gradient descent work with quadratically bounded noise?

Theorem

Yes. Converges at a $1/T$ *rate if objective is smooth and strongly convex, or* $1/\sqrt{T}$ *if smooth and merely convex. (D., Gairrigos, and Gower, 2023, Thms. 10+11)*

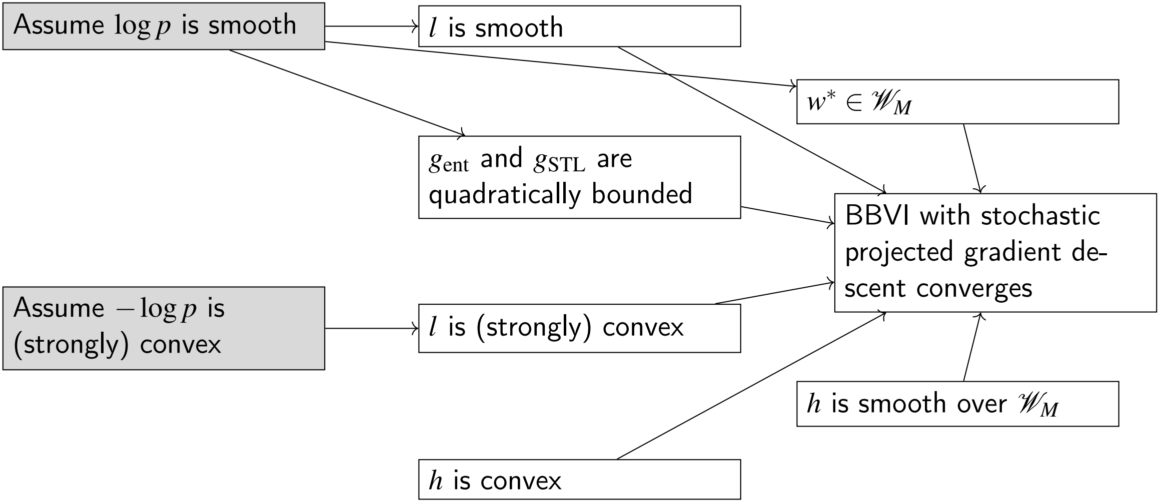
Putting the pieces together (proximal gradient descent)

Putting the pieces together (proximal gradient descent)



Putting the pieces together (projected gradient descent)

Putting the pieces together (projected gradient descent)



Putting the pieces together

Theorem

If $-\log p(z,x)$ *is* M*-smooth and (strongly) convex, then stochastic proximal gradient descent using the* g_{energy} *estimator with a dense Gaussian variational family with triangular* C *with an appropriate stepsize sequence converges to the optimum of the ELBO at a* $1/\sqrt{T}$ $(1/T)$ *rate. (D., Gairrigos, and Gower, 2023, Cor. 12)*

Putting the pieces together

Theorem

If $-\log p(z,x)$ is M-smooth and (strongly) convex, then stochastic proximal gradient descent using the g_{energy} estimator with a dense Gaussian variational family with triangular C with an appropriate stepsize sequence converges to the optimum of the ELBO at a $1/\sqrt{T}$ $(1/T)$ rate. (D., Gairrigos, and Gower, 2023, Cor. 12)

Theorem

If $-\log p(z,x)$ is M-smooth and (strongly) convex, then stochastic projected gradient descent (projecting onto $\mathscr{W}_M$) using either the g_{STL} or g_{ent} estimators with a dense Gaussian variational family with symmetric C with an appropriate stepsize sequence converges to the optimum of the ELBO at a $1/\sqrt{T}$ $(1/T)$ rate. (D., Gairrigos, and Gower, 2023, Cor. 13)

Outline

1. Introduction
2. The neg-entropy
3. The energy
4. Proximal gradient descent
5. Projected gradient descent
6. Gradient variance
7. Real convergence guarantees
8. Discussion

Related work

- Kim et al. (2023) give a similar $1/T$ rate for proximal SGD using g_{energy} with smoothness and strong convexity.

Related work

- Kim et al. (2023) give a similar $1/T$ rate for proximal SGD using g_{energy} with smoothness and strong convexity.
- Xu and Campbell (2023) give a $1/\sqrt{T}$ rate for projected-SGD using g_{ent} with a particular rescaling which is *asymptotic* in the number of observations (☹) and *local* (☹) but does not require convexity (☺).

Related work

- Kim et al. (2023) give a similar $1/T$ rate for proximal SGD using g_{energy} with smoothness and strong convexity.
- Xu and Campbell (2023) give a $1/\sqrt{T}$ rate for projected-SGD using g_{ent} with a particular rescaling which is *asymptotic* in the number of observations (☹) and *local* (☹) but does not require convexity (☺).
- Lambert et al. (2022) give a $1/T$ rate for a VI-like SGD algorithm from a discretization of a Wasserstein gradient flow with smoothness+strong convexity. Diao et al. (2023) give a related proximal with a $1/T$ rate or $1/\sqrt{T}$ with just convexity. These require the Hessian of the log-posterior (☹) but are very beautiful (☺).

Open questions

- Why does *regular* SGD seem to work so well?

Open questions

- Why does *regular* SGD seem to work so well?
- Guarantees with Adam instead of SGD?

Open questions

- Why does *regular* SGD seem to work so well?
- Guarantees with Adam instead of SGD?
- Guarantees without assuming we know smoothness/strong convexity constants?

Open questions

- Why does *regular* SGD seem to work so well?
- Guarantees with Adam instead of SGD?
- Guarantees without assuming we know smoothness/strong convexity constants?
- Guarantees without assuming smoothness or (strong) convexity at all?

Open questions

- Why does *regular* SGD seem to work so well?
- Guarantees with Adam instead of SGD?
- Guarantees without assuming we know smoothness/strong convexity constants?
- Guarantees without assuming smoothness or (strong) convexity at all?
- Guarantees with more general variational families (e.g. normalizing flows)?

Open questions

- Why does *regular* SGD seem to work so well?
- Guarantees with Adam instead of SGD?
- Guarantees without assuming we know smoothness/strong convexity constants?
- Guarantees without assuming smoothness or (strong) convexity at all?
- Guarantees with more general variational families (e.g. normalizing flows)?
- Is this "inference research" or "optimization research"?

Open questions

- Why does *regular* SGD seem to work so well?
- Guarantees with Adam instead of SGD?
- Guarantees without assuming we know smoothness/strong convexity constants?
- Guarantees without assuming smoothness or (strong) convexity at all?
- Guarantees with more general variational families (e.g. normalizing flows)?
- Is this "inference research" or "optimization research"?

Thank you!

these slides: `t.ly/sICHy` or `people.cs.umass.edu/domke/convergence.pdf`

Citations

- D. *Provable gradient variance guarantees for black-box variational inference*. NeurIPS 2019.
- D. *Provable smoothness guarantees for black-box variational inference.* ICML 2020.
- D., Gairrigos, and Gower. *Provable convergence guarantees for black-box variational inference.* NeurIPS 2023.
- Kim, Oh, Wu, Ma, and Gardner. *On the convergence and scale parameterizations of black-box variational inference.* NeurIPS 2023.
- Xu and Campbell. *The computational asymptotics of gaussian variational inference and the laplace approximation.* Stat Comput, (32), 2023.
- Lambert, Chewi, Bach, Bonnabel, and Rigollet. *Variational inference via Wasserstein gradient flows.* NeurIPS 2022.
- Diao, Balasubramanian, Chewi, and Salim. *Forward- backward Gaussian variational inference via JKO in the Bures-Wasserstein space.* ICML 2023.

Table of properties

$$F(w) := \underbrace{\mathbb{E}_{q_w(z)}[-\log p(z,x)]}_{\text{"energy" } l(w)} + \underbrace{\mathbb{E}_{q_w(z)} \log q_w(z)}_{\text{"neg-entropy" } h(w)}$$

Condition on $-\log p(z,x)$	**Consequence**
none	$h(w)$ convex (when C symmetric or triangular)
	$h(w)$ *not* strongly convex, *not* smooth
	$h(w)$ is M-smooth over $\mathscr{W}_M$
convex	$l(w)$ convex
c-strongly convex	$l(w)$ c-strongly convex
	$\|C\|_F^2 + \|m - z^*\|_2^2 \leq \frac{d}{c}$ at solution
M-smooth	$l(w)$ M-smooth
	$w^* \in \mathscr{W}_M$
	gradient estimators quadratically bounded