Stochastic Optimization

for Black-Box Variational Inference

Journées annuelles du GdR MOA - Université Perpignan Via Domitia

Guillaume Garrigos

October 2023

Université Paris Cité

A work in collaboration with

Justin Domke
University of Massachusetts

Robert M. Gower
Flatiron Institute

I : Introduction

Variational Inference

We have a distribution $p(x, z)$, where x is explicit data and z is latent variable

We want to estimate $p(z|x)$ with a simple family $\mathcal{Q} : p(\cdot|x) \sim q \in \mathcal{Q}$

Variational Inference

We have a distribution $p(x, z)$, where x is explicit data and z is latent variable

We want to estimate $p(z|x)$ with a simple family $\mathcal{Q} : p(\cdot|x) \sim q \in \mathcal{Q}$

$$\min_{q \in \mathcal{Q}} KL(q \,||\, p(\cdot|x)) = \int q(z) \ln \frac{q(z)}{p(z|w)} dz = \mathbb{E}_z \ \ln \frac{q(z)}{p(z|w)}$$

Variational Inference

We have a distribution $p(x, z)$, where x is explicit data and z is latent variable

We want to estimate $p(z|x)$ with a simple family $\mathcal{Q} : p(\cdot|x) \sim q \in \mathcal{Q}$

$$\min_{q \in \mathcal{Q}} KL(q \mid\mid p(\cdot|x)) = \int q(z) \ln \frac{q(z)}{p(z|w)} dz = \mathbb{E}_z \ln \frac{q(z)}{p(z|w)}$$

Equivalently

$$\min_{q \in \mathcal{Q}} \mathbb{E}_z \ln q(z) - \mathbb{E}_z \ln p(x, z) \tag{VI}$$

Variational Inference : Gaussian Family

$$\min_{q \in \mathcal{Q}} \ \mathbb{E}_z \ln q(z) - \mathbb{E}_z \ln p(x, z) \tag{VI}$$

Assumption (Gaussian Family)

We assume that $\mathcal{Q} = \{q_w \mid w \in \mathcal{W}^+\}$, with

- $\mathcal{W} = \mathbb{R}^d \times \mathcal{M}^d$ where $\mathcal{M}^d = \mathcal{T}^d$ (lower triangular) or $\mathcal{S}^d$ (symmetric)
- $\mathcal{W}^+ = \{(m, C) \in \mathcal{W} \mid C \succ 0\}$
- $q_w(z) = \mathcal{N}(z|m, CC^\top)$

Variational Inference : Gaussian Family

$$\min_{w \in \mathcal{W}^+} \mathbb{E}_z \ln q_w(z) - \mathbb{E}_z \ln p(x,z) \qquad \text{(VI)}$$

Assumption (Gaussian Family)

We assume that $\mathcal{Q} = \{q_w \mid w \in \mathcal{W}^+\}$, with

- $\mathcal{W} = \mathbb{R}^d \times \mathcal{M}^d$ where $\mathcal{M}^d = \mathcal{T}^d$ (lower triangular) or $\mathcal{S}^d$ (symmetric)
- $\mathcal{W}^+ = \{(m, C) \in \mathcal{W} \mid C \succ 0\}$
- $q_w(z) = \mathcal{N}(z|m, CC^\top)$

II : Structural properties

Properties of the entropy h

Proposition (Convexity of the entropy - Domke 2020)

Let $h(w) = \mathbb{E}_z \ln q_w(z) + \delta_{\mathcal{W}^+}(w)$.

1. $h(w) = -\ln\det C$ if $C \succ 0$, $+\infty$ otherwise
2. h is proper lower semi-continuous **convex** over $\mathcal{W} = \mathbb{R}^d \times \mathcal{M}^d$
3. $\mathrm{prox}_{\gamma h}(m, C) = (m, \hat{C})$ with $\hat{C}_{ii} \leftarrow \frac{1}{2}(C_{ii} + \sqrt{C_{ii}^2 + 4\gamma})$, if $\mathcal{M}^d = \mathcal{T}^d$

Proposition (Smoothness of the entropy - Domke 2020)

1. ∇h is L-**Lipschitz** over $\mathcal{W}_L^+ = \{(m, C) \in \mathcal{W}^+ \mid \sigma_{\mathrm{min}}(C) \geqslant \frac{1}{\sqrt{L}}\}$
2. $\mathrm{proj}_{\mathcal{W}_L^+}(m, C)$ can be computed by doing a SVD on C, if $\mathcal{M}^d = \mathcal{S}^d$

Properties of the free energy ℓ

Proposition (Convexity and smoothness of the energy - Domke 2020)

Let $\ell(w) = -\mathbb{E}_z \ln p(x, z)$.

1. If $-\ln p(\cdot, x)$ is convex then ℓ too
2. If $-\ln p(\cdot, x)$ is μ-strongly convex then ℓ too
3. If $-\ln p(\cdot, x)$ is L-smooth, then ℓ too
4. $\operatorname{argmin}(h + \ell) \subset \mathcal{W}_L^+ = \{(m, C) \in \mathcal{W}^+ \mid \sigma_{\min}(C) \geqslant \frac{1}{\sqrt{L}}\}$

Properties of the free energy ℓ

Proposition (Convexity and smoothness of the energy - Domke 2020)

Let $\ell(w) = -\mathbb{E}_z \ln p(x, z)$.

1. If $-\ln p(\cdot, x)$ is convex then ℓ too
2. If $-\ln p(\cdot, x)$ is μ-strongly convex then ℓ too
3. If $-\ln p(\cdot, x)$ is L-smooth, then ℓ too
4. $\operatorname{argmin}(h + \ell) \subset \mathcal{W}_L^+ = \{(m, C) \in \mathcal{W}^+ \mid \sigma_{\min}(C) \geqslant \frac{1}{\sqrt{L}}\}$

Assumption (log-concave and smooth target)

We assume that $-\ln p(\cdot, x)$ is convex and L-smooth

Properties of the free energy ℓ

Assumption (log-concave and smooth target)

We assume that $-\ln p(\cdot, x)$ is convex and L-smooth

Example (Models with log-concave and smooth target)

1. Bayesian linear regression
2. Logistic regression
3. Hierarchical logistic regression

Properties of the problem

$$\min_{w \in \mathcal{W}^+} \mathbb{E}_z \ln q_w(z) - \mathbb{E}_z \ln p(x,z) = h(w) + \ell(w) \tag{VI}$$

We can consider two approaches:

Properties of the problem

$$\min_{w \in \mathcal{W}^+} \mathbb{E}_z \ln q_w(z) - \mathbb{E}_z \ln p(x,z) = h(w) + \ell(w) \tag{VI}$$

We can consider two approaches:

1. h is prox-friendly, and ℓ is smooth : we do a **proximal stochastic gradient** method
   - encode with $\mathcal{M}^d = \mathcal{T}^d$ so that prox_h costs $O(d)$ operations

Properties of the problem

$$\min_{w \in \mathcal{W}^+} \mathbb{E}_z \ln q_w(z) - \mathbb{E}_z \ln p(x,z) = h(w) + \ell(w) \tag{VI}$$

We can consider two approaches:

1. h is prox-friendly, and ℓ is smooth : we do a **proximal stochastic gradient** method
   - encode with $\mathcal{M}^d = \mathcal{T}^d$ so that prox_h costs $O(d)$ operations
2. $f = h + \ell$ is smooth over $\mathcal{W}_L^+$: we do a **projected stochastic gradient** method
   - encode with $\mathcal{M}^d = \mathcal{S}^d$ so that $\text{proj}_{\mathcal{W}_L^+}$ is tractable $O(d^3)$

III : Stochastic algorithms

III : Stochastic algorithms

1 : Classical theory for SGD

To minimize $f(w) = \mathbb{E}_z f(w, z)$, the SGD algorithm writes

$$w^{t+1} = w^t - \gamma_t g^t, \quad \mathbb{E}_z \left[g^t\right] = \nabla f(w^t)$$

To minimize $f(w) = \mathbb{E}_z f(w, z)$, the SGD algorithm writes

$$w^{t+1} = w^t - \gamma_t g^t, \quad \mathbb{E}_z \left[g^t\right] = \nabla f(w^t)$$

Typical results in the convex setting are :

- $t^{-\frac{1}{2}}$ convergence when $\gamma_t \downarrow 0$: $\mathbb{E}\left[f(w^t) - \inf f\right] = O\left(\frac{1}{\sqrt{t}}\right)$
- ε^{-2} complexity when $\gamma_t \equiv \gamma$: $\mathbb{E}\left[f(w^t) - \inf f\right] = O\left(\frac{1}{\gamma t} + \gamma\sigma^2\right)$

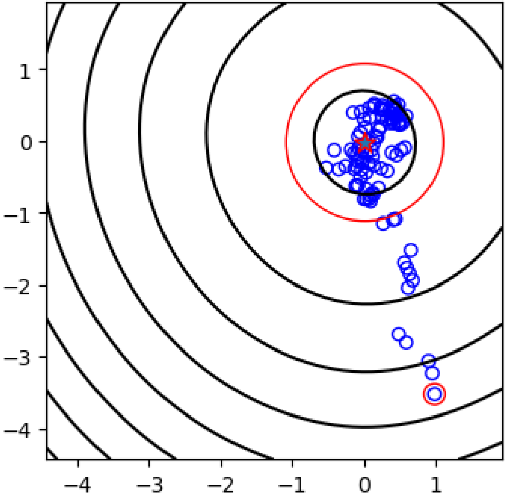
To minimize $f(w) = \mathbb{E}_z f(w,z)$, the SGD algorithm writes

$$w^{t+1} = w^t - \gamma_t g^t, \quad \mathbb{E}_z \left[g^t\right] = \nabla f(w^t)$$

Typical results in the convex setting are :

- $t^{-\frac{1}{2}}$ convergence when $\gamma_t \downarrow 0$: $\mathbb{E}\left[f(w^t) - \inf f\right] = O\left(\frac{1}{\sqrt{t}}\right)$
- ε^{-2} complexity when $\gamma_t \equiv \gamma$: $\mathbb{E}\left[f(w^t) - \inf f\right] = O\left(\frac{1}{\gamma t} + \gamma\sigma^2\right)$

Bonus : if no variance (interpolation holds) then we get *better* rates


1
0
−1
−2
−3
−4
−4
−3
−2
−1
0
1

To minimize $f(w) = \mathbb{E}_z f(w,z)$, the SGD algorithm writes

$$w^{t+1} = w^t - \gamma_t g^t, \quad \mathbb{E}_z \left[g^t\right] = \nabla f(w^t)$$

Typical results in the convex setting are :

- $t^{-\frac{1}{2}}$ convergence when $\gamma_t \downarrow 0$: $\mathbb{E}\left[f(w^t) - \inf f\right] = O\left(\frac{1}{\sqrt{t}}\right)$
- ε^{-2} complexity when $\gamma_t \equiv \gamma$: $\mathbb{E}\left[f(w^t) - \inf f\right] = O\left(\frac{1}{\gamma t} + \gamma\sigma^2\right)$

Usually require assumptions on f (regularity) and g^t (variance control):

- f is Lipschitz ☹ or ∇f is Lipschitz ☹ or $f(\cdot, z)$ is uniformly smooth ☹
- $\mathbb{E}_z\left[\|g^t\|^2\right] \leqslant C$ or $C\|\nabla f(w^t)\|^2$ ☹

To minimize $f(w) = \mathbb{E}_z f(w,z)$, the SGD algorithm writes

$$w^{t+1} = w^t - \gamma_t g^t, \quad \mathbb{E}_z \left[g^t\right] = \nabla f(w^t)$$

Typical results in the convex setting are :

- $t^{-\frac{1}{2}}$ convergence when $\gamma_t \downarrow 0$: $\mathbb{E}\left[f(w^t) - \inf f\right] = O\left(\frac{1}{\sqrt{t}}\right)$
- ε^{-2} complexity when $\gamma_t \equiv \gamma$: $\mathbb{E}\left[f(w^t) - \inf f\right] = O\left(\frac{1}{\gamma t} + \gamma\sigma^2\right)$

Usually require assumptions on f (regularity) and g^t (variance control):

- f is Lipschitz ☹ or ∇f is Lipschitz ☹ or $f(\cdot, z)$ is uniformly smooth ☹
- $\mathbb{E}_z\left[\|g^t\|^2\right] \leqslant C$ or $C\|\nabla f(w^t)\|^2$ ☹

We need **new optimization theory** for the **niche** properties verified by VI

III : Stochastic algorithms

2 : Proximal Stochastic Gradient method for VI

Remember $h(w) = \mathbb{E}_z \ \ln q_w(z) + \delta_{\mathcal{W}^+}(w)$, $\ell(w) = -\mathbb{E}_z \ \ln p(x,z)$

The **Proximal Stochastic Gradient Descent** method writes as:

$$w^{t+1} = \mathrm{prox}_{\gamma_t h}\left(w^t - \gamma_t g^t\right), \ \mathbb{E}\left[g^t\right] = \nabla \ell(w^t)$$

Remember $h(w) = \mathbb{E}_z \ln q_w(z) + \delta_{\mathcal{W}^+}(w)$, $\ell(w) = -\mathbb{E}_z \ln p(x, z)$

The **Proximal Stochastic Gradient Descent** method writes as:

$$w^{t+1} = \operatorname{prox}_{\gamma_t h}\left(w^t - \gamma_t g^t\right), \ \mathbb{E}\left[g^t\right] = \nabla \ell(w^t)$$

Lemma (The energy estimator)

If $u \sim \mathcal{N}(0, I)$ and $g^t_{\text{energy}} := -\nabla_w \ln p(x, C^t u + m^t)$, then

$$\mathbb{E}_u \left[g^t_{\text{energy}}\right] = \nabla \ell(w^t) \quad \text{and} \quad \mathbb{E}_u \left[\|g^t_{\text{energy}}\|^2\right] \leqslant A\|w - w^*\|^2 + B$$

The noise bound $O(\|w - w^*\|^2 + 1)$ is new, but we can exploit it to get rates

Remember $h(w) = \mathbb{E}_z \ \ln q_w(z) + \delta_{\mathcal{W}^+}(w)$, $\ell(w) = -\mathbb{E}_z \ \ln p(x,z)$

The **Proximal Stochastic Gradient Descent** method writes as:

$$w^{t+1} = \mathrm{prox}_{\gamma_t h}\left(w^t - \gamma_t g^t\right), \ \mathbb{E}\left[g^t\right] = \nabla \ell(w^t)$$

Theorem (Rates for solving VI)

Let w^t be generated by the above method, with the **energy** estimator g^t_{energy}.

1. for a suitable $\gamma_t \downarrow 0$, we have $\mathbb{E}\left[f(w^t) - \inf f\right] = O\left(\frac{1}{\sqrt{t}}\right)$
2. for a constant $\gamma_t \equiv \frac{1}{LT}$, we have $\mathbb{E}\left[f(w^T) - \inf f\right] = O\left(\frac{1}{\sqrt{T}}\right)$

Remember $h(w) = \mathbb{E}_z \ln q_w(z) + \delta_{\mathcal{W}^+}(w)$, $\ell(w) = -\mathbb{E}_z \ln p(x,z)$

The **Proximal Stochastic Gradient Descent** method writes as:

$$w^{t+1} = \operatorname{prox}_{\gamma_t h}\left(w^t - \gamma_t g^t\right), \ \mathbb{E}\left[g^t\right] = \nabla \ell(w^t)$$

Theorem (General optimization result)

Let ℓ be convex and L-smooth, let h be convex. Assume the estimator is quadratically bounded : $\mathbb{E}\left[\|g^t\|^2\right] \leqslant A\|w^t - w^*\|^2 + B$. If $\gamma \leqslant \frac{1}{L}$ then

$$\mathbb{E}\left[f(w^t) - \inf f\right] \simeq O\left(\frac{A}{\gamma t} + B\gamma\right)$$

Remember $h(w) = \mathbb{E}_z \ln q_w(z) + \delta_{\mathcal{W}^+}(w)$, $\ell(w) = -\mathbb{E}_z \ln p(x, z)$

The **Proximal Stochastic Gradient Descent** method writes as:

$$w^{t+1} = \mathrm{prox}_{\gamma_t h}\left(w^t - \gamma_t g^t\right), \; \mathbb{E}\left[g^t\right] = \nabla \ell(w^t)$$

Theorem (General optimization result)

Let ℓ be μ-convex and L-smooth, let h be convex. Assume the estimator is quadratically bounded : $\mathbb{E}\left[\|g^t\|^2\right] \leqslant A\|w^t - w^*\|^2 + B$. If $\gamma \leqslant \frac{1}{L}$ then

$$\mathbb{E}\left[f(w^t) - \inf f\right] \simeq O\left(A\theta_\gamma^t + B\gamma\right)$$

III : Stochastic algorithms

3 : Projected Stochastic Gradient for VI

Remember $h(w) = \mathbb{E}_z \ln q_w(z) + \delta_{\mathcal{W}^+}(w)$, $\ell(w) = -\mathbb{E}_z \ln p(x, z)$
ℓ is gloablly L-smooth, and h too over $\mathcal{W}_L^+ = \{w \in \mathcal{W}^+ \mid \sigma_{\min}(C)^2 \geqslant 1/L\}$

Our **Projected Stochastic Gradient Descent** method writes as:

$$w^{t+1} = \mathrm{proj}_{\mathcal{W}_L^+}\left(w^t - \gamma_t g^t\right), \; \mathbb{E}\left[g^t\right] = \nabla(\ell + h)(w^t)$$

Remember $h(w) = \mathbb{E}_z \ \ln q_w(z) + \delta_{\mathcal{W}^+}(w)$, $\ell(w) = -\mathbb{E}_z \ \ln p(x,z)$
ℓ is gloablly L-smooth, and h too over $\mathcal{W}_L^+ = \{w \in \mathcal{W}^+ \mid \sigma_{\min}(C)^2 \geqslant 1/L\}$

Our **Projected Stochastic Gradient Descent** method writes as:

$$w^{t+1} = \mathrm{proj}_{\mathcal{W}_L^+}\left(w^t - \gamma_t g^t\right), \ \mathbb{E}\left[g^t\right] = \nabla(\ell + h)(w^t)$$

Lemma (The entropy estimator)

If $u \sim \mathcal{N}(0, I)$ and $g^t_{\text{entropy}} := g^t_{\text{energy}} + \nabla h(w)$, then

$$\mathbb{E}_u\left[g^t_{\text{entropy}}\right] = \nabla f(w^t) \ \text{ and } \ \mathbb{E}_u\left[\|g^t_{\text{entropy}}\|^2\right] \leqslant A\|w - w^*\|^2 + B$$

Remember $h(w) = \mathbb{E}_z \ln q_w(z) + \delta_{\mathcal{W}^+}(w)$, $\ell(w) = -\mathbb{E}_z \ln p(x,z)$
ℓ is gloablly L-smooth, and h too over $\mathcal{W}_L^+ = \{w \in \mathcal{W}^+ \mid \sigma_{\min}(C)^2 \geqslant 1/L\}$

Our **Projected Stochastic Gradient Descent** method writes as:

$$w^{t+1} = \mathrm{proj}_{\mathcal{W}_L^+}\left(w^t - \gamma_t g^t\right), \ \mathbb{E}\left[g^t\right] = \nabla(\ell + h)(w^t)$$

Theorem (Rates for VI)

Let w^t be generated by the above method, with the **entropy** estimator g^t_{entropy}. For a suitable $\gamma_t \downarrow 0$ (or a constant $\gamma_t \equiv \frac{1}{LT}$), we have

$$\mathbb{E}\left[f(w^T) - \inf f\right] = O\left(\frac{1}{\sqrt{T}}\right)$$

Remember $h(w) = \mathbb{E}_z \ \ln q_w(z) + \delta_{\mathcal{W}^+}(w)$, $\ell(w) = -\mathbb{E}_z \ \ln p(x, z)$
ℓ is gloablly L-smooth, and h too over $\mathcal{W}_L^+ = \{w \in \mathcal{W}^+ \mid \sigma_{\min}(C)^2 \geqslant 1/L\}$

Our **Projected Stochastic Gradient Descent** method writes as:

$$w^{t+1} = \mathrm{proj}_{\mathcal{W}_L^+}\left(w^t - \gamma_t g^t\right), \ \mathbb{E}\left[g^t\right] = \nabla(\ell + h)(w^t)$$

Theorem (General optimization result)

Let $\ell + h$ be convex and differentiable on $\mathcal{W}_L^+$. Assume the estimator is quadratically bounded : $\mathbb{E}\left[\|g^t\|^2\right] \leqslant A\|w^t - w^*\|^2 + B$. If $\gamma \leqslant \frac{1}{L}$ then

$$\mathbb{E}\left[f(w^t) - \inf f\right] \simeq O\left(\frac{A}{\gamma t} + B\gamma\right)$$

Remember $h(w) = \mathbb{E}_z \ \ln q_w(z) + \delta_{\mathcal{W}^+}(w)$, $\ell(w) = -\mathbb{E}_z \ \ln p(x,z)$
ℓ is gloablly L-smooth, and h too over $\mathcal{W}_L^+ = \{w \in \mathcal{W}^+ \mid \sigma_{\min}(C)^2 \geqslant 1/L\}$

Our **Projected Stochastic Gradient Descent** method writes as:

$$w^{t+1} = \mathrm{proj}_{\mathcal{W}_L^+}\left(w^t - \gamma_t g^t\right), \ \mathbb{E}\left[g^t\right] = \nabla(\ell + h)(w^t)$$

Theorem (General optimization result)

Let $\ell + h$ be μ-convex and differentiable on $\mathcal{W}_L^+$. Assume the estimator is quadratically bounded : $\mathbb{E}\left[\|g^t\|^2\right] \leqslant A\|w^t - w^*\|^2 + B$. If $\gamma \leqslant \frac{1}{L}$ then

$$\mathbb{E}\left[f(w^t) - \inf f\right] \simeq O\left(A\theta_\gamma^t + B\gamma\right)$$

Remember $h(w) = \mathbb{E}_z \ \ln q_w(z) + \delta_{\mathcal{W}^+}(w)$, $\ell(w) = -\mathbb{E}_z \ \ln p(x, z)$
ℓ is gloablly L-smooth, and h too over $\mathcal{W}_L^+ = \{w \in \mathcal{W}^+ \mid \sigma_{\min}(C)^2 \geqslant 1/L\}$

Our **Projected Stochastic Gradient Descent** method writes as:

$$w^{t+1} = \mathrm{proj}_{\mathcal{W}_L^+}\left(w^t - \gamma_t g^t\right), \ \mathbb{E}\left[g^t\right] = \nabla(\ell + h)(w^t)$$

Lemma (The Stick The Landing (STL) estimator)

If $u \sim \mathcal{N}(0, I)$ and $g^t_{\mathrm{STL}} := g^t_{\mathrm{energy}} + \nabla_w \ln q_v(C^t u + m^t)$ with $v = w^t$, then

$$\mathbb{E}_u\left[g^t_{\mathrm{STL}}\right] = \nabla f(w^t) \ \text{ and } \ \mathbb{E}_u\left[\|g^t_{\mathrm{STL}}\|^2\right] \leqslant A\|w - w^*\|^2 + B$$

where $B = 0$ if the target distribution $p(\cdot|x)$ is a Gaussian.

Remember $h(w) = \mathbb{E}_z \ \ln q_w(z) + \delta_{\mathcal{W}^+}(w)$, $\ell(w) = -\mathbb{E}_z \ \ln p(x,z)$
ℓ is gloablly L-smooth, and h too over $\mathcal{W}_L^+ = \{w \in \mathcal{W}^+ \mid \sigma_{\min}(C)^2 \geqslant 1/L\}$

Our **Projected Stochastic Gradient Descent** method writes as:

$$w^{t+1} = \mathrm{proj}_{\mathcal{W}_L^+}\left(w^t - \gamma_t g^t\right), \ \mathbb{E}\left[g^t\right] = \nabla(\ell + h)(w^t)$$

Theorem (Exponential rates for VI with Gaussian target)

Let w^t be generated by the above method, with the **STL** estimator g^t_{STL}.
Assume that the target p is **Gaussian**. For a suitable γ_t, we have

$$\mathbb{E}\left[f(w^T) - \inf f\right] = O\left(\theta^T\right), \quad \theta \in [0,1).$$

IV : Conclusion

Conclusion and perspectives

- Black-box VI problems have very specific properties
  - estimator with quadratic noise $A\|w - w^*\|^2 + B$
  - non-global smoothness $\mathcal{W}_L^+$
- Required a new analysis of SGD
- Estimate how well STL works when target is Gaussian
  - What if the target is almost Gaussian?
- In practice people do SGD without projection on $\mathcal{W}_L^+$: is this needed at all?
- Can we get results without convexity but Polyak-Łojasiewicz? (we tried)

Thank you for your attention !